

基于主题挖掘采煤工作面事故致因模型探索



孟凡强^{1,*}, 李春霞²

¹北京赛迪工业和信息化工程监理中心工程公司, 北京 100089

²泰山职业技术学院机电技术工程系, 山东泰安 271001

摘要: 采煤工作面作为煤炭第一生产现场, 是煤矿事故多发地。为避免或减少事故的发生, 需要及时排除发现采煤工作面事故隐患, 但以往事故致因多采用定性分析为主的方法, 缺少有效的事故致因分析模型, 以至于大量事故隐患数据潜在的价值没有得到应用。为此, 采用 LDA 主题挖掘算法, 构建事故致因主题挖掘模型, 采用困惑度(Perplexity)指标来确定最佳主题个数, 并利用语义网络图分析事故隐患之间的内在关系, 直观地反映了采煤工作面主要事故致因及致因因素间的关联。以山东某矿 2013 年 8 月~2019 年 9 月 18925 条事故隐患数据为例, 通过对主题模型挖掘, 抽取事故致因主题, 并绘制语义网络图, 应用社会网络分析方法分析事故隐患要素之间的关联关系。结果表明: 管理缺陷、人员不安全行为、设备不安全状态及环境不安全状态是导致采煤工作面事故的主要因素, 它们之间的相互作用导致了事故的发生。

关键词: 采煤工作面; 事故隐患; 事故致因; 主题挖掘; 社会网络分析

DOI: [10.57237/j.ssrf.2022.01.006](https://doi.org/10.57237/j.ssrf.2022.01.006)

Research on Cause of Coal Face Accident Based on Topic Model Mining

Meng Fanqiang^{1,*}, Li Chunxia²

¹Beijing CCID Industry & Information Engineering Supervision Center Co., Ltd, Beijing 100089, China

²Department of Mechanical and Electrical Engineering, Taishan Polytechnic, Tai'an 271001, China

Abstract: As the first production site of coal, mining working face is the place where coal mine accidents occur frequently. In order to avoid or reduce the occurrence of accidents, it is necessary to eliminate and discover the hidden danger of accidents in mining face in time. However, in the past, the cause of accidents mainly adopts the method of qualitative analysis, and lacks the effective analysis model of the cause of accidents, so that the potential value of a large number of hidden danger data of accidents has not been applied. Therefore, the LDA topic mining algorithm is used to build the accident cause topic mining model, and the Perplexity index is used to determine the best number of topics, and the semantic network graph is used to analyze the internal relationship between the hidden dangers of accidents, which directly reflects the main accident cause and the correlation between the cause factors of coal mining face. Taking 18925 accident hidden danger data of a mine in Shandong Province from August 2013 to September 2019 as an example, the topic of accident causes was extracted by mining the topic model, and the semantic network diagram was drawn. The

*通信作者: 孟凡强, meng_fq@163.com

social network analysis method was applied to analyze the association relationship between the factors of accident hidden danger. The results show that management defects, unsafe behavior of personnel, unsafe state of equipment and unsafe state of environment are the main factors leading to coal face accidents, and their interaction leads to the occurrence of accidents.

Keywords: Coal Working Face; Hidden Danger of Accident; Cause of Accident; Theme Mining; Social Network Analysis

1 引言

采煤工作面生产系统是煤矿生产重要子系统，也是煤矿事故的高发区。采煤生产系统的安全生产事故指在生产过程中，使生产系统短暂或较长时间或永久中止运行，或出现人员伤亡或物损失的事故[1]。由于煤矿生产的特殊性，采煤事故也具有独特性，直接用于采煤生产系统的事故致因理论比较缺乏，而且又因为采煤生产系统的特殊性及事故因素的复杂性，需要建立科学、合理的事故致因的理论，用以调查、分析、评价煤矿事故[2]。

事故致因理论是通过分析大量安全事故发生原因，挖掘事故的本质原因而提炼出事故机理模型，比较典型的理论有事故频发倾向理论、事故因果连锁理论、轨迹交叉理论、能量意外释放理论及系统安全理论等[3]。

2 文献研究

梳理煤矿事故统计分析、煤矿事故致因理论相关的文献资料，事故致因研究可归纳为以下几个方面：

- (1) 事故地域特征：根据全国煤矿灾害事故统计分析，通过研究煤矿灾害类型特征、事故类型、事故等级、事故地域特性等方面[4]，研究我国不同区域煤矿事故灾害类型，研究结果表明南方省份发生事故频数比北方省份高[5]，而且重特大事故发生次数多。
- (2) 事故时间特征：在全国煤矿事故统计分析基础上，研究发现煤矿灾害事故发生的高发时间是春季、夏季，尤其集中在7、8月份，每天的交接班是事故发生的高发时间段[6, 7]。
- (3) 事故灾害类型：通过研究，瓦斯、水灾与顶板是导致煤矿事故主要灾害因素[8]，煤矿企业要重点防控瓦斯爆炸、顶板管理、水害防治。
- (4) 事故本质致因因素：煤矿事故本质致因因素包

括直接原因、间接原因、本质原因与信息原因。直接原因为人的不安全行为、物的不安全状态、不安全的环境；间接原因包括技术原因、人员素质原因、人员健康原因、作业环境影响等[9]；本质原因中存在3种观点，管理失误论、利润驱使论和安全文化缺失论；信息原因指安全管理、安全预警系统建设与完善程度、应急管理体系建设等[10]。

上述研究现是从宏观角度研究全国煤矿事故致因机理，对于采煤工作面生产系统而言，有着独自的特征，其事故机理也有独自的特殊性，需要结合采煤工作面特点进行研究。

3 采煤工作面事故致因主题挖掘

根据以往煤矿事故历史统计数据，通过定性分析而获得采煤工作面事故致因因素，但由于记录的文本数据只是对特、重大事故进行比较详细的描述，而对绝大多数的轻伤或无伤亡的事故，只是对事情的过程、原因做下简单记录，诸如此类记录数量远远比详细的事故调查报告数量多。虽然记录文本粗糙、文字简单，但蕴涵着事故前后经过、作业环境和人员状况等信息，对分析事故致因具有极大的利用价值。需要应用数据挖掘技术，从大量文本数据中提取关键的信息，进一步挖掘出有价值的信息，找出影响采煤工作事故发生的致因因素[11]。

主题模型（Topic Model）是在自然语言应用领域中，从一系列文档中抽象主题的一种统计方法[12]，而潜在狄利克雷分配（Latent Dirichlet Allocation, LDA）[13]是一种全概率生成模型，按照一定规则从海量文本中抽取主题，再根据主题分别对文本进行聚类的数据挖掘算法模型。

LDA主题模型结构模型如图1所示。

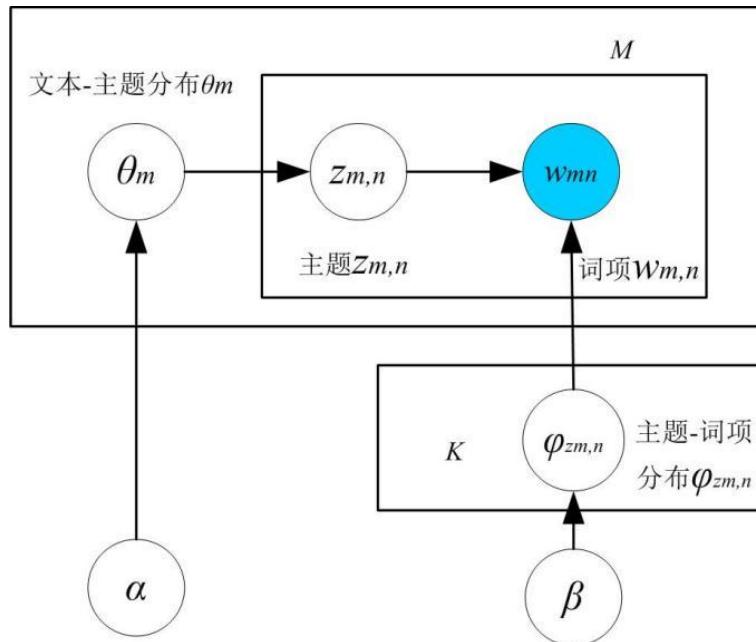


图 1 LDA 主题模型结构

图中, α 、 β 表示狄利克雷先验参数, 即语料库参数; K 代表主题总数, M 代表文本集合 N 中文本总数; m 是 M 中的一个文本, θ_m 表示 m 的概率分布, 它是 k 维向量; n 指的是文本 m 中的词项数, 第 m 个文本的主题全体记为 z_m , 而 $z_{m,n}$ 指的是第 n 个词项对应的主题;

第 m 个文本中的所有特征词记为 w_m , 其中, $w_{m,n}$ 表示第 m 个文本中的第 n 个特征词; $\varphi_{zm,n}$ 代表主题的词项分布。 θ_m 、 $\varphi_{zm,n}$ 都服从狄利克雷分布 (Dirichlet Distribution); $z_{m,n}$ 和 $w_{m,n}$ 服从多项分布[13]。

给定参数 α 与 β , 假设 M 份文本中, 各文本之间的主题相互独立, 通过已知的某个文本的主题条件概率分布后, 可得到文本集合中的所有文本的主题生成概率, 求得 LDA 模型中的联合分布概率以及后验概率分布如下所示。

$$\begin{cases} p(\theta_{1:N}, \phi_{1:k}, z_{1:N}, w_{1:N}) = \\ \prod_{i=1}^K p(\phi_i) \prod_{m=1}^N p(\theta_m) \left(\prod_{n=1}^M p(z_{m,n} | \theta_m) \right) p(w_{m,n} | \phi_{1:k}, z_{m,n}) \quad (1) \\ p(\theta_{1:N}, \phi_{1:k}, z_{1:N} | w_{1:N}) = \frac{p(\theta_{1:N}, \phi_{1:k}, z_{1:N}, w_{1:N})}{p(w_{1:N})} \end{cases}$$

在 LDA 主题模型中, 需要求解主题-文档 θ_m 概率分布和主题-词项 $\varphi_{zm,n}$ 的概率分布。常用的求解方法主要有 Gibbs 抽样算法和 VEM 变分算法, 由于在实际应

用中, Gibbs 抽样算法操作简单、易于实现[14], 所以选用该方法进行参数估计。

求得 θ_m 和 $\varphi_{zm,n}$ 计算公式, 如下所示。

$$\begin{cases} \theta_m = \frac{n_m^{(k)} + \alpha_k}{\sum_{k=1}^K n_m^{(k)} + \alpha} \\ \varphi_{zm,n} = \frac{n_k^{(i)} + \beta_i}{\sum_{i=1}^M n_k^{(i)} + \beta_i} \end{cases} \quad (2)$$

4 应用案例

4.1 数据来源与预处理

以山东某矿采煤工作面为研究对象, 通过现场调研获取 2013 年 8 月~2019 年 9 月安全监测监控系统和生产安全调度管理系统自动采集和人工排查隐患记录, 共计 18925 条。为了保证文本挖掘得到好的效果, 先对原文本数据数据进行清洗、初步规范化整理, 只保留隐患排查日期、隐患内容、隐患排查地点和责任单位。

4.2 事故致因主题挖掘

本文采用困惑度 (Perplexity) 指标来确定最佳主题个数, 一般而言, 困惑度越小, 则模型的性能越好[15]。经计算, 本案例最优主题数为 5。

利用 R 语言中软件包对收集到的数据进行处理, 获得相应的事故数据分词。使用 ROSTCM6 对分词进行处理, 获得高频词, 并依据 LDA 主题模型实现事故

致因主题抽取、语义网络构建。

建立特征词及对应特征向量权重(见表 1)。特征词的权重值越高, 则表示该词在文档中的重要性。

表 1 部分特征词及特征向量权重矩阵

序号	特征词	权重	序号	特征词	权重	序号	特征词	权重
1	安全	0.114	13	教育	0.003	25	原因	0.039
2	爆炸	0.069	14	井下	0.220	26	运输	0.075
3	当班	0.047	15	齐全	0.012	27	证照	0.011
4	导致	0.016	16	煤矿	0.486	28	职工	0.003
5	地方	0.039	17	人员	0.026	29	重庆	0.066
6	地区	0.037	18	事故	0.499	30	作业	0.064
7	顶板	0.247	19	水灾	0.032	31	运输	0.016
8	工区	0.016	20	死亡	0.476	32	皮带	0.013
9	管理	0.004	21	四川	0.077
10	贵州	0.039	22	突出	0.043
11	国有	0.042	23	乡镇	0.293
12	湖南	0.066	24	意识	0.003

利用 ROSTCM6 中 NetDraw 绘制特征词社会关系网络图, 事故文本集合中, 重要程度越高的特征词越靠近网络中心位置[16], 如图 2 所示。图中, “管理”、“事故”、“安全”、“作业”等处于中心位置, 周边是与其有关系的特征词, 通过带有箭头的直线连接。如中心词“安全”与“防护”、“检查”、“落实”、“制度”、“规程”、“执行”、“操作”、“违章”等词直接关联, 说明在防护落实、制度和规程执行、操作违章等方面存在影响安全的因素。

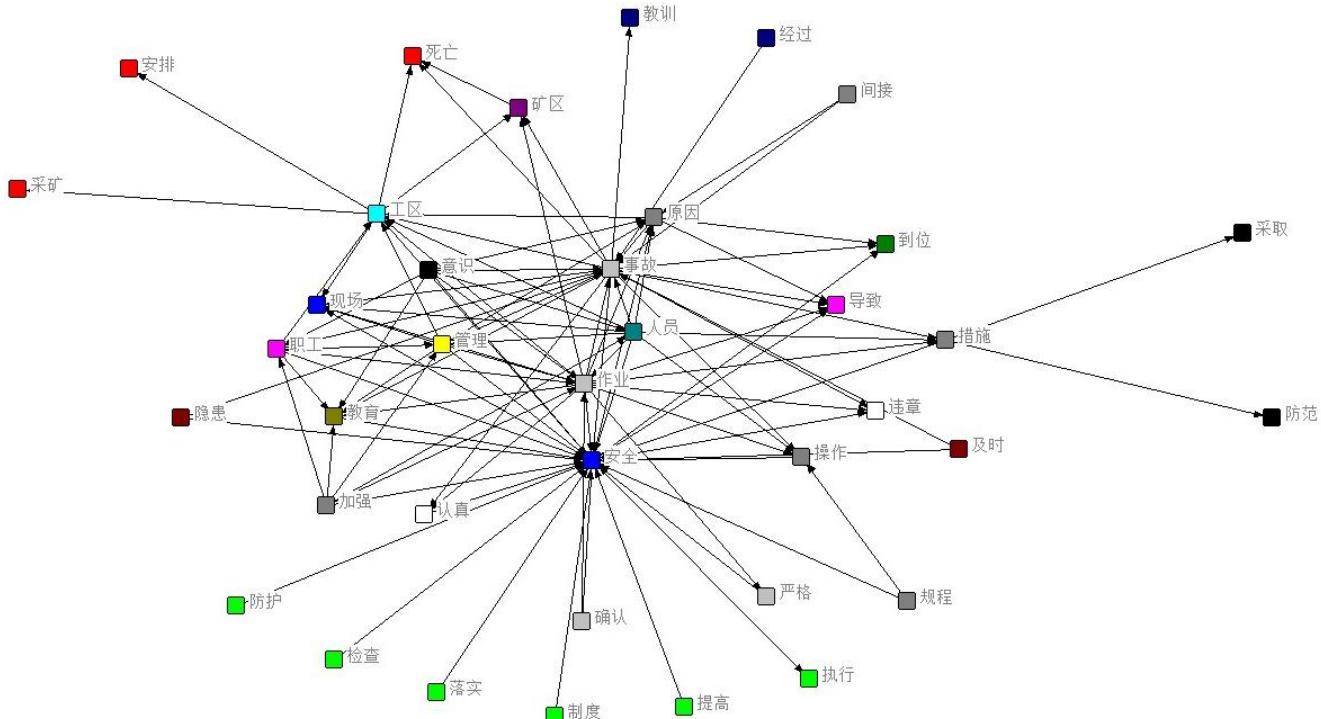


图 2 特征词关系网络图

对主题模型求解后, 得到事故致因主题及特征词, 事故致因主题抽取结果如下表所示。

表 2 事故致因主题挖掘结果

事故致因主题	主题词
管理主题	不严、加强、法规、制度、规程、措施、落实、执行、培训、不够、安全、教育、不到位、现场、工区
人员主题	意识、违章、措施、操作、缺乏、不够、淡薄、规程、不力
设备设施主题	防护、检查、维护、违章、完好、缺陷、故障、作业、操作
作业环境主题	顶板、冒落、瓦斯、火灾、突出、爆炸、事故、水灾
生产安全主题	原因、人员、作业、到位、导致、意识、管理、违章、提高、操作、认真、加强、现场

4.3 结果分析

综合表 1、表 2 及图 2，分析导致采煤工作面事故安全的原因：

- (1) 管理主题。由表 2 可知，导致事故发生的管理方面原因是：安全管理规章制度执行不到位，安全教育培训要加强，现场作业安全规程执行不严格，以及安全措施不够。由图 2 看到，“现场”、“工区”两个词处于中心位置，说明它们在网络中重要程度较高，表明作业现场多发事故，与工区管理不到位有直接关系。
- (2) 人员主题。由表 2 可知，导致事故发生的人员方面存在安全意识缺乏、违章操作、安全措施不到位等。结合图 2 可知，人员主题和管理主题在特征词方面有着不少重叠，表明正是由于管理不到位，才导致现场作业安全意识淡薄。
- (3) 设备设施主题。设备设施主题主要和“防护”、“检查”、“维护”、“违章”等特征词分析。现场设备设施安全防护有缺陷、检查不到位、违章操作，以及维护不及时，导致出现隐患，引起事故的发生。由表 1 看到，出现“运输”、“皮带”等特征词，说明在皮带运输方面存在隐患，导致事故的发生。
- (4) 作业环境主题。由表 1 可知，有“顶板”、“瓦斯”、“水灾”、“火灾”、“突出”、“爆炸”等特征词，表明在现场作业环境中，存在冒顶、水害、火灾、瓦斯突出、瓦斯爆炸等隐患，严重影响着采煤工作面安全。
- (5) 生产安全主题。由图 2 及表 2 可知，生产安全主题和以上几个主题有很多交叉共用的特征词，正是因为管理、人员存在不安全行为，设备设施、作业环境存在不安全状态，才影响生产安全，导致事故的发生。

5 结论

通过主题挖掘算法，以某煤矿 2013 年 8 月~2019 年 9 月事故隐患文本数据为样本，通过挖掘关联词之间的关系，抽取到 5 个与采煤工作面事故致因相关的主题，分别为管理主题、人员主题、设备设施主题、作业环境主题、生产安全主题。进一步借助社会网络分析法，分析与各主题关联的特征词，通过图表直观地反映了影响采煤工作面的主要因素，即采煤工作面事故发生是管理缺陷、人的不安全行为、设备不安全状态、环境的不安全状态相互作用而造成的。该方法改变了以往研究依靠定性分析的不足，为分析采煤工作面事故致因提供了思路，具有一定的参考价值。

参考文献

- [1] Meng Fanqiang. Safety Warning Model of Coal Face Based on FCM Fuzzy Clustering and GA-BP Neural Network [J]. Symmetry, 2021, 13 (6): 1082.
- [2] 张东, 聂百胜, 王龙康. 我国煤矿安全生产事故的致灾因素分析 [J]. 中国安全生产科学技术, 2013, 9 (5): 136-140.
- [3] Kai W, Jiang S, Zhang W, et al. Study on PCPR Security System Construction in Coal Mine [J]. Procedia Engineering, 2011, 26: 2044-2050.
- [4] 诸利一, 吕文生, 杨鹏, 等. 2007—2016 年全国煤矿事故统计及发生规律研究 [J]. 煤矿安全, 2018, v. 49; No. 528 (07): 237-240.
- [5] 孙玲, 宫立昊. 2019 年国内煤矿安全事故统计分析及对策研究 [J]. 决策探索 (中), 2020, No. 642 (02): 22-23.
- [6] 陈娟, 赵耀江. 近十年来我国煤矿事故统计分析及启示 [J]. 煤炭工程, 2012 (03): 145-147.
- [7] 许多康. 国有煤矿事故统计分析与防控对策 [J]. 山东煤炭科技, 2019 (12): 190-193.
- [8] 李波, 巨广刚, 王珂, 等. 2005—2014 年我国煤矿灾害事故特征及规律研究 [J]. 矿业安全与环保, v. 43; No. 227 (03): 111-114.

- [9] 康国峰. 煤矿安全隐患排查与治理分级分类管理机制探索 [J]. 煤矿安全, 2011, 42 (010): 158-160.
- [10] 王丹, 刘琳, 张小曼. 灰色关联度法在煤矿本质安全评价中的改进及应用 [J]. 中国安全生产科学技术, 2013, 9 (001): 151-157.
- [11] Meng, Fanqiang and Li, Chunxia. Safety Warning of Coal Mining Face Based on Big Data Association Rule Mining [J]. Journal of Computational Methods in Sciences and Engineering, 2022, 22 (4): 1035-1052.
- [12] 夏春艳. 数据挖掘技术与应用 [M]. 冶金工业出版社, 2014.
- [13] 牛毅, 樊运晓, 高远. 基于数据挖掘的化工生产事故致因主题抽取 [J]. 中国安全生产科学技术, 2019, 15 (10): 165-170.
- [14] 徐守坤, 周佳, 李宁, 等. 基于 word2vec 和 LDA 的文本主题 [J]. 计算机工程与设计, 2018, 39 (9): 2764-2769.
- [15] 何天文, 王红. 基于语义语法分析的中文语句困惑度评价 [J]. 计算机应用研究, 2017, 34 (12): 3638-3546.
- [16] 谭章禄, 王兆刚, 胡翰, 姜萱, 彭胜男. 基于文本聚类的煤矿安全隐患类型挖掘研究 [J]. 中国安全科学学报, 2019, 29 (3): 145-148.